Analysis and Implementation of Rotation-invariant Neural Network Architectures for Feature Extraction*

VINCENT RESS[1], MARKUS BRAENDLE[2] & NORBERT HAALA[1]

When aligning images from different domains, such as thermal (IR) and human-visible light (RGB) images, classical feature extraction methods such as SIFT or SURF encounter severe limitations. While new techniques utilizing CNNs enable corresponding assignments, their inherent non-equivariance to rotations restricts possible areas of applications. Within our work adaptions to the network architecture and the trainings pipeline to achieve a rotation-equivariant behaviour are discussed. The D2-Net, which is based on the VGG16 architecture and gains remarkable performance especially with regard to changes of the image domain, was used as reference. Through analyses on the HPatches dataset, significantly improved equivariance properties were achieved for all adaptation types investigated.

1 Introduction

While many applications in photogrammetry and computer vision would not have been feasible without classic feature extraction methods such as Scale-Invariant Feature Transform (SIFT) (LOWE 2004) or Speeded-Up Robust Features (SURF) (BAY et al. 2006), new methods based on neural networks are becoming increasingly important. For example, an increasing number of city councils are using thermal cameras to carry out aerial surveys to determine the heating requirements and to evaluate energy saving potentials of their local municipalities. Usually, these IR images are captured and co-registered in combination with images in the human visible range of light (RGB) to collect further 3D geodata such as building types or structures. To achieve high registration accuracies, it is helpful to extract and match features directly from both image domains. Since classic feature extraction methods reach their limits with these kind of tasks, new approaches are required. Comparable limitations can be seen when combining rendered images based on digital building models or Computer Aided Design (CAD) models with images of the real objects, as used in automated quality control, for example (GHIMIRE et al. 2021). SATTLER et al. (2018) were able to demonstrate that with a sufficient data basis, new approaches based on Convolutional Neural Networks (CNN)s outperform conventional feature extraction algorithms, especially in the case of changes in the image domain, image areas with low texture or larger differences in the perspective view (SATTLER et al. 2018).
Due to the sharing of weights within 'receptive fields' CNNs are equivariant to translations of the image content. This is an important characteristic for applications in the field of pattern recognition or feature extraction, as it allows objects or interest points to be detected and matched regardless of their position in the image. However, since CNNs do not have a 'natural' equivariance to

[1] Universität Stuttgart, Institut für Photogrammetrie und Geoinformatik, Geschwister-Scholl-Straße 24D, D-70174 Stuttgart, E-Mail: [vincent.ress, norbert.haala]@ifp.uni-stuttgart.de
[2] MBDA Deutschland GmbH, Hagenauer Forst 27, D-86529 Schrobenhausen, E-Mail: markus.braendle@mbda-systems.de
*This work was supported by MBDA Deutschland GmbH

 DOI: 10.24407/KXP:1885284888

rotations, many applications require a predefined orientation of the images for a reliable extraction and description of keypoints or other features. Although diverse training methods and network architectures to achieve rotation equivariance have been developed and evaluated for tasks like pattern recognition and image segmentation, there is a lack of comparable assessments for deep architectures in the field of feature extraction.
As our main contribution various methods that achieved promising results in the field of pattern recognition were implemented and evaluated within our work to increase the tolerance towards rotated input data for deep CNNs. With the aim of extracting rotationally equivariant features, adaptions of the training pipeline as well as changes in the network architecture were examined. The D2-Net (DUSMANU et al. 2019) architecture, which gains competitive performance especially in regard to changes of the image domain, was used as reference.
In the following chapters the most promising methods analysed in this work are described (Chapter 2), the implemented network adoptions and evaluation routines explained (Chapter 3) and the results of the examinations presented (Chapter 4).

2 Related Work and Selected Approaches

While a large number of publications presented rotation-equivariant network architectures or network layers such as Group equivariant Convolution Neural Networks (G-CNN) (COHEN & WELLING 2016), Harmonic Networks (H-Net) (WORALL et al. 2017) or Rotation Equivariant Vector Field Networks (RotEqNet) (Marcos et al. 2017), the evaluations were mostly targeted for problems in the field of pattern recognition or image segmentation. The presence of various fully-connected layers in the provided network architectures impedes the seamless transferability of the results to feature extraction tasks. The authors, MARCOS et al. (2017) conducted the analysis of correct assignment rates in comparison to the SIFT extractor on the Notre Dame dataset (WINDER & BROWN 2007), extracting descriptors through a shallow Siamese network. Approaches such as LF-Net (ONO et al. 2018) are estimating a dense orientation map within the network to use the additional orientation information for the keypoint selection. This architecture increases the tolerance towards rotated input images, but requires a rotation augmented training of the network. Furthermore, none of the related publications carried out the impact of equivariant network layers on the inference times. The following section introduces the most promising approaches for the development of rotation-equivariant network architectures.

2.1 Vector Field Networks (RotEqNet)

The idea of Vector Field Networks is based on the masking and rotation of the filter kernels learned during the training process and their application to the input data. To reduce the memory requirements during the forward pass through the network, the resulting feature maps from the rotation are combined into 2D vector fields through orientation pooling. The magnitude of the vector resulting from a pixel position is determined by the largest response of the filter at that location (compared across all feature maps of a filter kernel), and the direction is determined by the associated rotation angle. In the subsequent layers, two filter kernels are trained for the x- and y-coordinates of the vectors during the training process. Since the rotation of the filter kernels

occurs at discrete angles predetermined in the network design, it is not an exact method for achieving rotation equivariance. The investigation of the RotEqNet was extended within this thesis as it exhibited the lowest classification error on the rotated MNIST dataset (LECUN et al. 1998) compared to the other network adaptations. (MARCOS et al. 2017)

2.2 Harmonic Networks (H-Net)

The Harmonics Networks published by WORALL et al. (2017) are using so called 'kernel constraints' to constrain the shape of the filter kernels within the layers to the circular harmonics (cf. Eq. 1)

$$W_m(r, \phi; R, \beta) = R(r)e^{i(m\phi+\beta)}$$

Equation 1: Circular Harmonics within the complex plane

The coefficients r and ϕ are representing the image coordinates in polar form, $m \in Z$ is called rotation order, $R: R \rightarrowtail R^+$ defines the shape of the respective filter and is called the radial profile and $\beta \in [0,2\pi)$ represents the phase offset. During the training, the radial profile R(r) and the phase offset β will be optimized. The cross-correlation ($\star$) of an image $F^0 := F(r, \varphi)$ and the cross-correlation of the identically origin-rotated image $F^\theta := F(r, \pi^\theta[\phi])$ with $W_m := W_m(r, \phi; R, \beta)$ is described by the relation presented in Equation 2.

$$[W_m \star F^\theta] = e^{im\theta}[W_m \star F^0]$$

Equation 2: Relation of the cross-correlation of the images F^0 and F^θ

In this context, it becomes apparent that the equivariance behaviour is depending on the rotation order m, and theoretically, full equivariance can be achieved when $m = 1$. The equivariance behaviour of a layer results from a linear combination of the equivariance properties of the corresponding filter kernels (WORALL et al. 2017).

3 Modifications to the Selected Approaches (Methods)

All adjustments to the training pipeline and network architecture presented in this chapter are based on the D2-Net reference architecture and pipeline published by DUSMANU et al. (2019) For the introduction of H-Net layers, the E2CNN framework (WEILER & CESA 2021) was applied, which facilitates the representation of complex filters in the polar plane through the use of 'irregular representations'. The implementation of the RotEqNet layers is mainly based on the published source code of the authors (MARCOS et al. 2017) and has been adapted accordingly.
The methods presented in the following sub-chapters are divided into adaptations to the trainings pipeline such as data augmentation, and adaptations to the network architecture through the introduction of H-Net and RotEqNet layers. In the subsequent tests, all three methods were analysed separately.

3.1 Training and Network Adaptions

The adaptations to the training pipeline were implemented through a rotation-augmented training. For this purpose, the training data was rotated around the principal point of the image by discrete angles of 5° up to ±180°. To improve the training performance and create uniformly sized input

data, matching areas of the provided high-resolution images of the dataset were cropped before the hand-over to the encoder of the D2-Net.
To achieve an equivariant network architecture, the layers of the D2-Net where substituted by layers of the H-Net and the RotEqNet. In a first step the equivariance of the untrained adapted network architectures was analysed by measurements of the mean feature map distance of the last layer of the encoder. For this purpose, a non-rotated image and multiple by steps of five-degrees rotated images were passed through the randomly initialized network. The resulting feature maps were re-transformed and the mean distance of the corresponding feature vectors from the feature map of the rotated to the feature map of the unrotated image was calculated. Through parameter studies such as adjusting kernel sizes, changing pooling methods or modifying the activation functions the feature map distance was minimized and the equivariance behaviour optimized.
As a compromise between the number of operations and the proportion of masked weights, the size of the filter kernels of the network architecture adopted by RotEqNet-Layer was increased to 5×5 pixels. Based on the results of the parameter studies the vector fields were layerwise transformed back to scalar feature maps by discarding the information of the orientation.
For the network architecture adopted by layer of the H-Net the rotation order m of the corresponding filter kernels was based on the results of WEILER & CESA (2021) limited to $m \in [0,1]$. The use of higher rotation orders increases the computational effort and memory consumption without significantly improving the equivariance behaviour of the network. To compensate the lower number of training weights caused by the restriction to circular harmonics, the number of input and output channels per layer were doubled in comparison to the original D2-Net architecture. Similar to the RotEqNet layers, the best equivariance properties were achieved in the parameter studies by a layerwise back-transformation from complex to scalar feature maps.
The weights of the adopted network architectures were tuned with randomly selected data of the MegaDepth dataset (LI & SNAVELY 2018) which consists of 150k tourist images of landmarks worldwide (Fig. 1). The camera poses and the depth information of the images – required for the loss assessment - were calculated using the COLMAP Structure-from-Motion pipeline (SCHOENBERGER & FRAHM 2016).

3.2 Evaluation

The equivariance of the trained models was evaluated on rotated images of the MegaDepth dataset (Fig. 1) (LI & SNAVELY 2018). The extracted descriptors of the rotated image were matched with the descriptors of a non-rotated reference image and the corresponding keypoints and the associated keypoints were transformed back to the reference. Subsequently the mean matching accuracy (MMA) with respect to the angle of rotation was calculated. The MMA describes the ratio of correctly matched feature points to the number of all (nearest neighbour) matched feature points and is a robust measure for the quality of a feature extraction method and the uniqueness of the resulting feature descriptors. A point is considered as correctly matched, if the Euclidian distance from a projected point of a transformed image to a matched point of a reference image is below a certain threshold (radius).

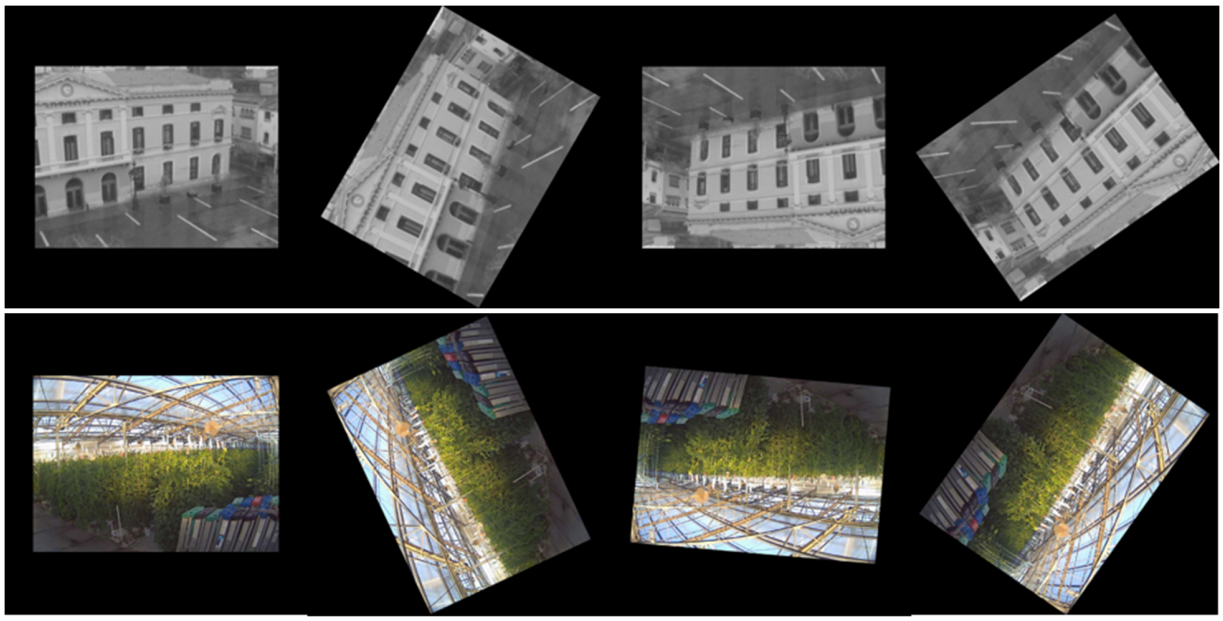

Fig. 1: Rotated sample images of the MegaDepth dataset

The evaluation of the general extraction performance was based on the HPatches benchmark. The benchmark dataset consists of 116 reference images modified by multiple photometric and perspective transformations (Fig. 2). Through the introduction of restricted random transformations (noise) different grades of difficulty are achieved. The CNNs to be analysed were used to extract feature points and descriptors. The known transformations from the reference images to the modified images were used to calculate the MMA of the matched feature points. During the final evaluation, the results are differentiated depending on the underlying type of transformation (photometric/perspective) (BALNTAS et al. 2017).
In addition to the analysis on the benchmark, the relative mean inference times in comparison to the reference architecture were determined.

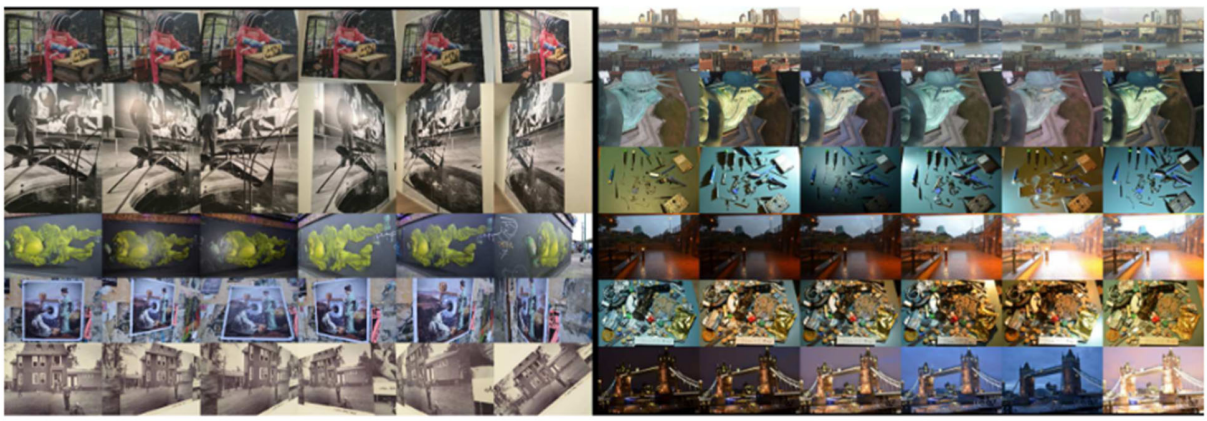

Fig. 2: Sample images of the HPatches dataset

4 Results

Fig. 3 demonstrates the impact of the adapted network and training structure on the equivariance behaviour of the D2-Net architecture. The graph clearly illustrates that all approaches show a remarkable improvement in handling rotated input data. Among them, the H-Net layer (green line) adaptation achieves the best results, exhibiting a matching accuracy that remains nearly consistent regardless of the angle of rotation. The augmented trained D2-Net (dashed blue line) model tends to form local minima especially at low thresholds in the angular ranges of 90° and 270°. The introduction of RotEqNet-Layer (red line) leads to local maxima for the angles of 90°, 180° and 270°, with the determined MMA dropping by up to 30 percentage points in the angular ranges in between.

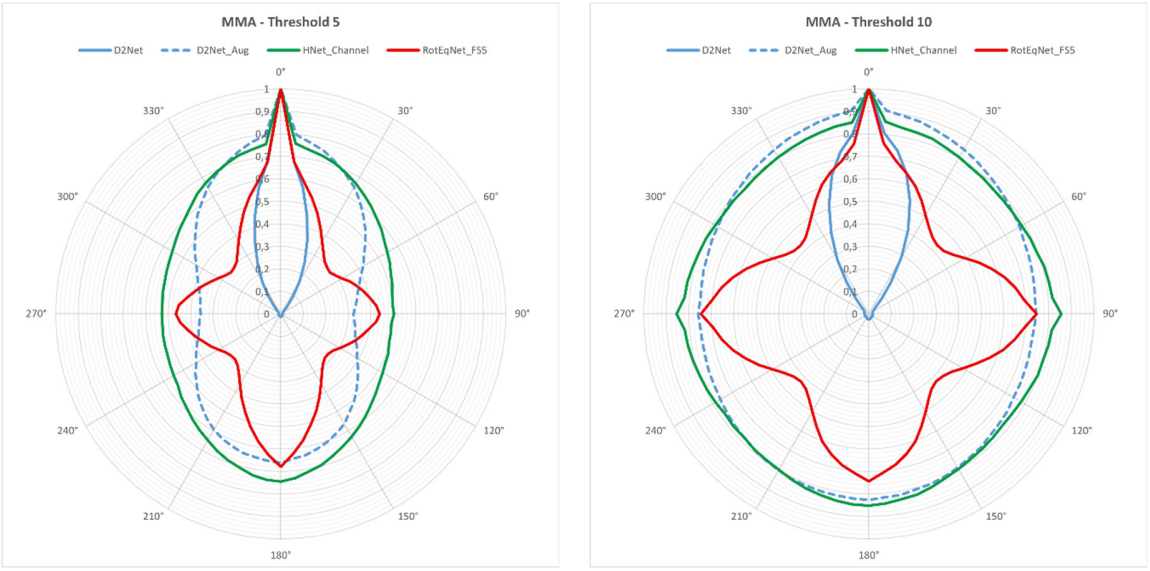

Fig. 3: Representation of the Mean Matching Accuracy (MMA) (radial axis) as a function of the angle of rotation (angle) for the analysed approaches

For the CNNs evaluated on the HPatches benchmark the resulting MMAs (y-axis) are visualized for the thresholds from 1 pixel to 10 pixels (x-axis) (Fig. 4). The best overall result (Overall) on the HPatches benchmark is achieved by the augmented trained D2-Net architecture. While particularly with regard to perspective transformations (Fig. 4 – Viewpoint) the results of the reference architecture are surpassed, for photometric transformations (Fig. 4 – Illumination) only small improvements for thresholds below 4 pixels are noticeable. By using equivariant H-Net layers, comparable MMAs are achieved up to a threshold of 4 pixels to the reference architecture. However, these drop significantly for larger thresholds below those of the D2-Net. By far the lowest overall performance of all types of adaptation considered is obtained by the introduction of RotEqNet layers.

The measured values shown in Table 1 illustrate that the modification of the D2-Net architecture with H-Net layers raises inference times by 280%, while the adaptation with RotEqNet layers results in 3000% increased inference times.

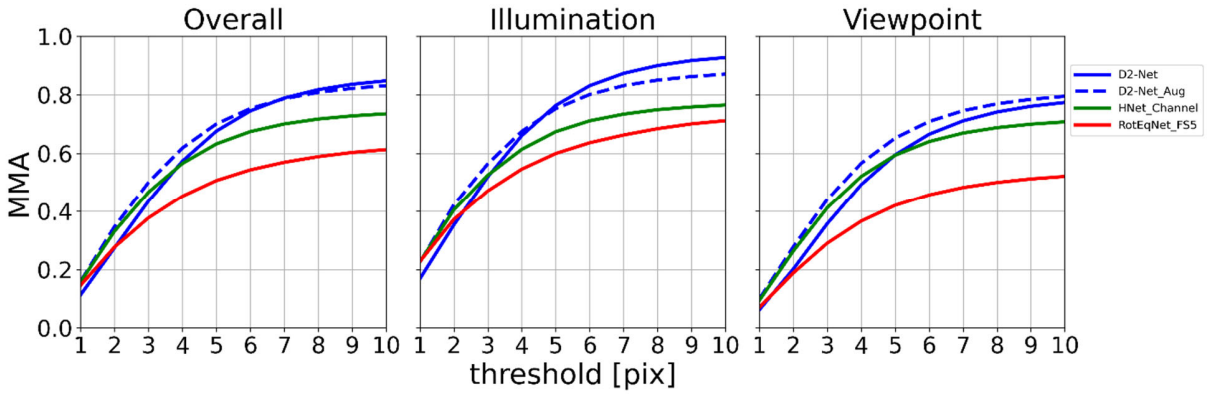

Fig. 4: Results of the HPatches benchmark analysis

Table 1: Measurements of the trained net weights and the relative inference times

	D2-Net	D2-Net_Aug	HNet_Channel	RotEqNet_FS5
no. of weights	7.6M	7.6M	6.8M	21.2M
rel. inference times [%]	100	102	277	3021

5 Discussion and Conclusion

The surprisingly poor overall performance of the RotEqNet layers on the HPatches benchmark can be explained by the masking and interpolation method used by the authors for the rotation of the filter kernels. The resulting proportion of non-masked weights depends on the filter size defined in the design and, for the small filter sizes used, leads to strongly limited development potential of the filter weights. The high inference times are caused by additional convolution operations for the rotated filter kernels. Modifying the masking and interpolation methods employed by the authors could potentially enhance overall performance on the HPatches benchmark for smaller kernel sizes. However, such adjustments are unlikely to positively impact computational efficiency.

The results of adaptations by the H-Net layers are in line with expectations from previous publications, both in terms of equivariance behaviour and general feature extraction properties. The difference in overall performance on the HPatches dataset can be explained by the "kernel constraint" and the associated limited development possibilities of the filter kernels.

The augmented network outperforms the reference network by a few percentage points, especially for higher accuracy requirements with thresholds from 2 pixels to 6 pixels. This suggests a high percentage of unused weights in the reference architecture that are ‘activated’ by the augmentation, thus preserving the range of different filter kernels specialized for the task. However, well-founded statements in this regard can only be made on the basis of further analyses of the filter structures.

In summary, the equivariance behaviour of the reference architecture could be significantly improved on the basis of all types of adaptation presented, allowing a robust tie point matching based on CNN-based approaches also on rotated imagery. Especially the training on shallow network architectures and on small datasets may lead to improved performance by the introduction of equivariant layer types such as H-Net or RotEqNet. These results will be surpassed by the augmented trained network architectures, if sufficiently large datasets are available for training.

6 Literature

BALNTAS, V., LENC, K., VEDALDI, A. & MIKOLAJCYK, K., 2017: HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. IEEE Conference on Computer Vision and Pattern Recognition, http://arxiv.org/pdf/1704.05939v1.

BAY, H., TUYTELAARS, T. & VAN GOOL, L., 2006: SURF: Speeded Up Robust Features. Computer vision - ECCV, **3951**, Lecture Notes in Computer Science, 404-417.

COHEN, T. S. & WELLING, M., 2016: Group Equivariant Convolutional Networks, Proceedings of the International Conference on Machine Learning (ICML), https://arxiv.org/pdf/1602.07576.

DUSMANU, M., ROCCO, I. & PAJDLA, T., 2019: D2-Net: A Trainable CNN for Joint Detection and Description of Local Features. IEEE Conference on Computer Vision and Pattern Recognition, https://arxiv.org/pdf/1905.03561.

GHIRMIRE, P., JOVANCEVIC, I. & ORTEU, J.-J., 2021: Learning Local Descriptor for Comparing Renders with Real Images. Appl. Sci., https://doi.org/10.3390/app11083301.

LI, Z. & SNAVELY, N., 2018: MegaDepth: Learning Single-View Depth Prediction from Internet Photos. IEEE Conference on Computer Vision and Pattern Recognition, http://arxiv.org/pdf/1804.00607v4.

LECUN, Y., BOTTOU, L., BENGIO, Y. & HAFFNER, P., 1998: Gradient-based learning applied to document recognition. Proceedings of the IEEE, **86**, 2278-2324.

LOWE, D. G., 2004: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, **60**(2), 91-110.

MARCOS, D., VOLPI, M. & KOMODAKIS, N., 2017: Rotation Equivariant Vector Field Networks. IEEE International Conference on Computer Vision (ICCV), https://arxiv.org/pdf/1612.09346.

ONO, Y., TRULLS, E., FUA, P. & MOO YI, K., 2018: LF-Net: Learning Local Features from Images, Advances in Neural Information Processing Systems, **31**, (NeurIPS 2018), http://arxiv.org/pdf/1805.09662.

SATTLER, T., MADDERN, W., TOFT, C., HAMMARSTRAND, L., STENBORG, E., SAFARI, D., OKUTOMI, M., POLLEFEYS, M., SIVIC, J., KAHL, F. & PAJDLA, T., 2018: Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions. IEEE Conference on Computer Vision and Pattern Recognition, https://arxiv.org/pdf/1707.09092.pdf.

SCHOENBERGER, J. L. & FRAHM, J.-M., 2016: Structure-from-Motion Revisited. 29th IEEE Conference on Computer Vision and Pattern Recognition, 4104-4113, https://demuc.de/papers/schoenberger2016sfm.pdf, letzter Zugriff 03.02.24.

WEILER, M. & CESA, G. 2021: General E(2)-Equivariant Steerable CNNs, Proceedings of Conference on Neural Information Processing Systems (NeurIPS), https://arxiv.org/pdf/1911.08251.

WINDER, S. A. J. & BROWN, M., 2007: Learning Local Image Descriptors, 2007 IEEE Conference on Computer Vision and Pattern Recognition, 1-8.

WORALL, E. D., GARBIN, J. S., TURMUKHAMBETOV, D. & BROSTOW, G. J., 2017: Harmonic Networks: Deep Translation and Rotation Equivariance. IEEE Conference on Computer Vision and Pattern Recognition, https://arxiv.org/pdf/1612.04642.